

Gene integrated set profile analysis: a context-based approach for inferring biological endpoints

Jeanne Kowalski^{1,2,*}, Bhakti Dwivedi², Scott Newman², Jeffery M. Switchenko^{1,2}, Rini Pauly¹, David A. Gutman³, Jyoti Arora¹, Khanjan Gandhi⁴, Kylie Ainslie², Gregory Doho⁵, Zhaohui Qin^{2,3}, Carlos S. Moreno^{1,6}, Michael R. Rossi^{1,7}, Paula M. Vertino^{1,7}, Sagar Lonial^{1,8}, Leon Bernal-Mizrachi^{1,8} and Lawrence H. Boise^{1,8}

¹Winship Cancer Institute, Emory University, Atlanta, GA 30333, USA, ²Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, Atlanta, GA 30333, USA, ³Department of Biomedical Informatics and Neurology, School of Medicine, Emory University, Atlanta, GA 30322, USA, ⁴Department of Human Genetics, School of Medicine, Emory University, Atlanta, GA 30322, USA, ⁵Centers for Disease Control, Atlanta, GA 30322, USA, ⁶Department of Pathology and Laboratory Medicine, School of Medicine, Emory University, Atlanta, GA 30322, USA, ⁷Department of Radiation Oncology, School of Medicine, Emory University, Atlanta, GA 30322, USA and ⁸Department of Hematology and Medical Oncology, School of Medicine, Emory University, Atlanta, GA 30322, USA

Received July 14, 2015; Revised November 05, 2015; Accepted December 10, 2015

ABSTRACT

The identification of genes with specific patterns of change (e.g. down-regulated and methylated) as phenotype drivers or samples with similar profiles for a given gene set as drivers of clinical outcome, requires the integration of several genomic data types for which an ‘integrate by intersection’ (IBI) approach is often applied. In this approach, results from separate analyses of each data type are intersected, which has the limitation of a smaller intersection with more data types. We introduce a new method, GISPA (Gene Integrated Set Profile Analysis) for integrated genomic analysis and its variation, SISPA (Sample Integrated Set Profile Analysis) for defining respective genes and samples with the context of similar, a priori specified molecular profiles. With GISPA, the user defines a molecular profile that is compared among several classes and obtains ranked gene sets that satisfy the profile as drivers of each class. With SISPA, the user defines a gene set that satisfies a profile and obtains sample groups of profile activity. Our results from applying GISPA to human multiple myeloma (MM) cell lines contained genes of known profiles and importance, along with several novel targets, and their further SISPA application to MM coMMpass trial data showed clinical relevance.

INTRODUCTION

The widespread availability of cancer genomics data prompts a critical, yet unanswered question: How do we identify genetic drivers of a particular phenotype given the diverse ways in which genes can be dysregulated? For a single data type, such as gene expression (GE), well-established methods exist, but introducing additional data types such as CpG methylation, copy number (CN) and somatic gene mutations make an integrated analysis much more challenging. Assuming that genetic drivers can be identified from multidimensional data, the subsequent question of how to identify ‘similar’ samples within another data set then arises. When a single data type is present, methods of assessing similarity exist (1,2), but finding ‘similar’ samples among multidimensional data is a considerable challenge.

Integrate by intersection (IBI) is the simplest approach to analyzing multidimensional data of several data types. With IBI, the results of independent analyses from each data type are intersected post hoc. While easily implemented, a major limitation of this approach is that as the number of data types increases, the intersection becomes small and smaller. Various modeling approaches have also been used for integrated analyses, which often require large sample sizes and generally assume an analytical distribution describing the relationship among data types (3). For example, in examining differential methylation associated GE changes, one assumes that expression is modulated by differential methylation.

Gene Integrated Set Profile Analysis (GISPA) is a novel approach that combines and compares several genome-

*To whom correspondence should be addressed. Tel: +404 778 5305; Fax: +404 778 5016; Email: Jeanne.kowalski@emory.edu

wide data types from three or more sample classes in order to find the drivers of each class. GISPA produces ranked gene sets within the context of an a priori specified molecular profile, such as genes that have some combination of increased CpG methylation, CN loss and decreased GE specific to a single sample or class. Sample Integrated Set Profile Analysis (SISPA), a variation of GISPA, is a novel approach to find samples within the context of a similar, a priori multidimensional profile from a gene set of interest, either GISPA-defined or by the user. GISPA and SISPA derive results from a combined analysis of all data types; both are non-parametric and therefore do not rely upon imposed analytical distributions and crucially, do not require a large sample size.

Here, we apply GISPA to RNA-Seq, DNA CpG methylation and DNA CN data from three, extensively studied human multiple myeloma (MM) cell lines: KMS11 (4), MM1s (5) and RPMI8226 (6). Having identified potential driving genes' profiles specific to each cell line, we apply SISPA to identify patients with similar driving gene profiles from a large MM clinical trial. Finally, we derive a differential prognostic, mutation dependency network based on GISPA-defined sample-specific mutation profiles.

MATERIALS AND METHODS

Materials

Data generation. DNA and RNA were isolated from human myeloma cell lines and applied to array-based platforms: Illumina Omni1 Quad and Illumina Infinium Human Methylation 450K following the manufacturers' protocols. For RNA-Seq, 3 μ g of total RNA was obtained using the Illumina HiSeq at \sim 1000X coverage. Prior to analysis, proportions (e.g. CpG methylation beta values, variant proportions) were transformed using $\log_2((1 + p) / (1 - p))$, and GE data were transformed using $\log_2(\text{DESeq} + 1)$. All microarray and RNA-Seq data analyses were done based on (RefSeq) annotated, non-pseudo genes located on chromosomes 1 thru 22. Details on data processing are contained in the Supplement.

Clinical associations. Data were obtained from the ongoing Multiple Myeloma Research Foundation (MMRF) CoMMpass Trial (NCT0145429), a longitudinal study in MM relating clinical outcomes to genomic and immunophenotypic profiles of CD138 selected plasma cells from the bone marrow of newly diagnosed MM patients (7). Data from 377 patients with available clinical outcomes, Exome-Seq somatic mutations and CN segments and RNA-Seq ensemble GE at pre-treatment were downloaded based on the IA6 release of this trial from the MMRF researcher gateway portal (<https://research.themmr.org>). Data were similarly transformed as with the HMCL's. Sample z-scores and gene set were obtained using the bioconductor R package, 'GSVA' (2) and a change point model applied to a constructed composite score to define sample classes with and without profile activity for a given gene set using our SISPA method variation. Overall and progression-free survival analyses were done using SAS 9.3 (SAS Institute, Inc., Cary, NC, USA).

Network pathway analysis. The functional interaction (FI) network analysis was done using the Reactome FI approach (8) and HyperModules software plugins (9), as implemented in Cytoscape v3.2 (10).

miRNA associations. An in-house merger database (<http://mirnamerger.org/>) was used to search for known or predicted miRNA-mRNA target pairs against several public databases for a gene set.

Methods

Herein, we adopt The Cancer Genome Atlas (TCGA) nomenclature (<https://tcga-data.nci.nih.gov/tcga/dataAccessMatrix.htm>) that references a specific data type (RNA-Seq expression, DNA CpG methylation, etc.) as a feature, with specific sub features such as a gene, probe or CpG site collectively referred to as loci. A single-feature profile is defined as a combination of locus-level changes within a feature, while a combination of locus-level changes summarized across features is used to define a multi-feature profile.

Overview of GISPA. GISPA considers experiments with genome-wide data on one or more features from samples belonging to three or more classes. Given a user-defined molecular profile, GISPA derives ranked gene lists that satisfy this profile specific to each class. Below, we describe the four fundamental steps of GISPA (Figure 1), with the details provided in Supplementary Methods.

Step 1: specifying a profile. A profile is defined by specifying *a priori*, a change of either increase or decrease within each of the features. For example, eight possible profiles exist for three features. An example of a three feature profile relevant to cancer that we highlight in the results is that of decreased GE with increased CpG methylation and decreased CN.

Step 2: calculating within-feature profile statistics (WFPS). We introduce a novel statistic to filter genome-wide profile changes as drivers for each class. Assume, for example, three (single-sample) classes, C1, C2 and C3 in which we characterize C1 relative to C2 and C3 on the basis of our specified profile. We define gene sets that satisfy this profile as drivers of C1 based on a filter statistic (see Supplementary Methods) with the following properties: (i) $C1 - C2 \gg 0$; (ii) $C1 - C3 \gg 0$; and (iii) $C2 - C3 \approx 0$. Altogether, these properties minimize the potential for ambiguities in defining gene sets. Low WFPS values more closely correlate with the profile, whereas high values correlate with an extreme opposite profile. An empirical cumulative distribution function is constructed among the WFPS to define percentiles (see Supplementary Methods, Eq. 3). Using the WFPS percentiles, we define a between-genes, within-feature profile statistic (BG-WFPS, see Supplementary Methods). The use of percentiles enables a standardization of data such that the range between 0 and 1 is the same for each feature.

Step 3: calculating between-features profile statistic (BFPS). We construct a between feature profile statistic by summing

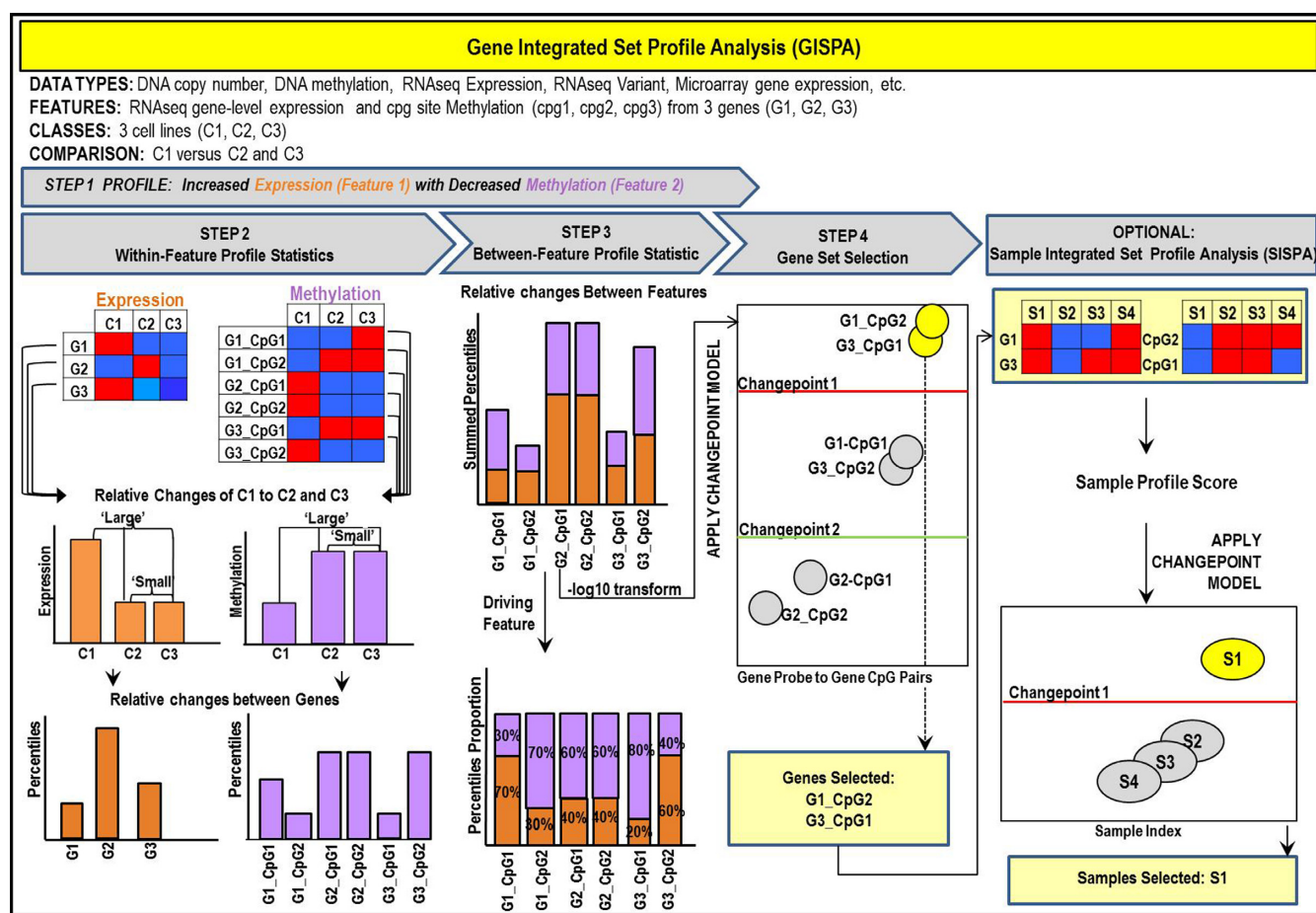


Figure 1. A Gene Integrated Set Profile Analysis (GISPA) overview illustrating the method. A three gene (G1, G2, G3) data set of two features, RNA-Seq gene expression (GE; feature 1) and CpG methylation of two sites (feature 2) as an example from three cell lines (C1, C2, C3). C1 is the cell line of interest to be compared against C2 and C3 based on the specified profile of increased GE and decreased CpG methylation (step 1). Using normalized data, shown as grids for each feature with 'high' (in red) and 'low' (in blue) values, a profile statistic is defined for each gene, within each feature among the cell lines that is based on the differences between C1 with C2 and C3, and between C2 and C3. This statistic is then used to obtain percentiles (step 2) that are summed between features for all probe and CpG site combinations to define a between feature profile statistic (step 3). Based on the summed percentiles, a prominent feature driving the between-feature profile statistic for each gene is defined as the feature with the maximum percentage contribution to the summed percentiles (step 4). This percent contribution assigns smaller values with a higher percentage than large values, since small values of the summed percentiles imply greater profile support. A multiple change point model (cpm) is applied to the $(-\log_{10})$ transformed summed percentiles to select genes with extreme values (change point 1) followed by the next most extreme (change point 2), etc. and denote these gene sets as ranked levels of support for the specified profile (step 4). Using these ranked gene sets and patient data, a sample profile score is defined based on a within-gene z-score among samples that is summed among genes within each feature, and in this example, subtracted between features, and applied in a second cpm to define sample classes with and without profile activity (see Supplementary Methods).

the WFPS percentiles for each gene. For some data types, such as CpG methylation, the number of CpG sites varies greatly among genes and therefore genes with a large number of CpG sites are more likely to be selected as part of a gene set. To address this issue, we generate a BFPS statistic for each gene and CpG site combination, and select the CpG site with the smallest summed percentile to be carried forward in step 3. In the case of transcript level GE data, we construct a combined data set based on compatible feature combinations such that if a CpG site does not correspond to a transcript, their combination is filtered. We obtain each feature's percentage of the summed percentiles and use this to define the feature associated with the highest percentage as the prominent feature. Since low values of the WFPS correlate with the profile of interest, we define an inverse proportion such that lower values reflect a higher percentage

contribution to the total than high values (see Supplementary Methods).

Step 4: deriving gene lists. We derive ranked gene sets according to various levels of support for the molecular profile of interest by applying a change point model (cpm) to the transformed $(-\log_{10})$ summed percentiles (BFPS). The use of a cpm in this context is analogous to applying a segmentation algorithm to array CN data, in which segments are typically defined by a breakpoint so that neighboring regions have different mean intensities. With our method, gene sets are formed by change points that are defined by successive differences in variances in the distribution of transformed percentiles. Because of the interpretation of our profile statistic, these gene sets are ranked according to those that most satisfy the profile (change point

1), next most (change point 2), etc. Additionally, one is able to estimate the statistical significance of genes sets in characterizing a class by ‘gene randomization’ (see Supplementary Methods).

SISPA. SISPA defines samples with and without profile activity, on average, among genes defined by a set, by applying a cpm to a composite, between features z-score formed by adding or subtracting individual sample scores between features, depending on the profile, between features (see Figure 1 and Supplementary Methods).

The R code to run GISPA is available for download at <https://github.com/BhaktiDwivedi>. The Bioconductor R package for SISPA is available at <https://www.bioconductor.org/packages/release/bioc/html/SISPA.html>.

RESULTS

All gene set results for two- and three-feature profiles are displayed in Supplementary Tables S1–S4.

Single-feature GISPA

As a proof of principle, we applied GISPA to single features, firstly SNP array CN and then RNA-Seq GE from the KMS11, MM1s and RPMI8226 MM cell lines (see Methods, Appendices 2 and 3). GISPA identified a deleted gene set specific to KMS11 that contained *TP53* (Supplementary Figure S1), and a set specific to MM1s containing *CDKN2A* (data not shown); these gene deletions are known abnormalities in the three cell lines (4,5). GISPA additionally identified *FGFR3* with increased GE specific to KMS11. This cell line has a t(4;14) translocation that results in *FGFR3* over expression. Interestingly, *FGFR3* fell between the first (cpt1) and second change points (cpt2). For RNA-Seq GE data in particular, there is often a mixture of ‘non-zero’ and ‘zero’ GE data such that genes with non-zero GE in the class of interest and zero GE in the comparison classes are first selected in cpt1 and genes with non-zero GE in all classes in cpt2. In such cases, the extension to more than one cpt may be needed to accommodate such a mixture, as in the case of *FGFR3*, a gene with non-zero GE in all three classes that is specifically up-regulated KMS11. Thus, if using a single feature profile of increased GE, one may consider using gene sets from three change points.

Two-feature GISPA

We next generated two-feature profiles incorporating decreased RNA-Seq GE with decreased CN. Among the KMS11 selected genes (Figure 2; Supplementary Figure S2 for MM1s and RPMI cell lines), *TP53* was, again, the topmost, showing CN change as the prominent feature. Among the other selected genes, *MAML2* had not previously been associated with MM and was also found as part of a novel gene fusion (11). Additionally, we applied GISPA to identify RNA-Seq derived coding variants with increased GE (Supplementary Figure S3), and coding variants with increased CpG methylation (Supplementary Figure S4) by implementing a ‘carry one forward’ approach (see Supplementary Methods) due to the varying number of

variants and CpG sites per gene. Variant status was used as a continuous variable based around the mutant allele fraction predicted from RNA-Seq so that even low-level mosaic mutations are considered in the analysis. GISPA identified known and novel expressed variants specific to each cell line. Among the GISPA defined genes supporting the profile of variants with increased GE in the KMS11 cell line as compared to both MM1s and RPMI8226 cell lines (a.k.a. ‘KMS11 selected expressed variants’), *FGFR3* was the top gene showing, as expected, a previously reported missense Y373C mutation (COSMIC ID, COSM718) with highly skewed GE toward the mutated copy (all RNA-Seq reads contained the variant allele; Supplementary Table S2) (12).

Three-feature GISPA

We next used GISPA to identify genes that satisfied the three-feature, cancer relevant profile of having loci with decreased GE and CN with increased CpG methylation as cell line-specific drivers. KMS11 selected genes with this profile are shown in Figure 3 (Supplementary Figure S5 for MM1s and RPMI cell lines).

We compared GISPA to the intersecting approach for defining KMS11 selected genes with our three-feature profile. We identified genes with increased CpG methylation in KMS11 versus MM1s and RPMI cell lines using a one-sided *t*-test and derived a list of differentially expressed genes using the DESeq package (13). By applying a *P*-value threshold of 0.01 for both methylation and GE results, a GE fold-change of at least 2, and requiring a CN loss specific to KMS11, no genes were identified as intersecting among the data type results. After removing the 2-fold GE fold-change criterion and applying the same *P*-value threshold, 40 genes were identified, of which *MERTK*, was in common with the GISPA gene set. These 40 genes were mostly associated with variable methylation changes which the GISPA profile penalizes by the additional requirement of little to no methylation differences between comparative cell lines.

MERTK was among the top KMS11 selected genes with a profile of decreased GE and decreased CN with increased CpG methylation as compared to both MM1s and RPMI cell lines with profiles of increased GE with decreased CpG methylation and no CN change (Figure 3; Supplementary Table S4). Additionally, a CN change is shown as the prominent feature driving this profile in *MERTK* for KMS11 (Figure 3). *MERTK* is a suspected oncogene, over-expressed in many different cancers and has been found to be mutated in MM and melanoma leading to a potential oncogenic function (14,15).

Among the other KMS11 selected genes with decreased GE and increased CpG methylation with decreased focal CN changes (CN segment less than 2% of the chromosome arm), *VILL* and *TTC22* show GE as the prominent feature, while *EPS15* (Early Growth Factor Pathway Substrate 15) show CpG methylation (Figure 3). *EPS15* is involved in translocations in acute leukemia (16), but as yet, has not been linked to MM. Whole arm or interstitial deletions of the chromosome 1p are observed in approximately 30% of myeloma patients and are associated with a poor prognosis (17–19). Deletions of 1p12 and 1p32.3 are of par-

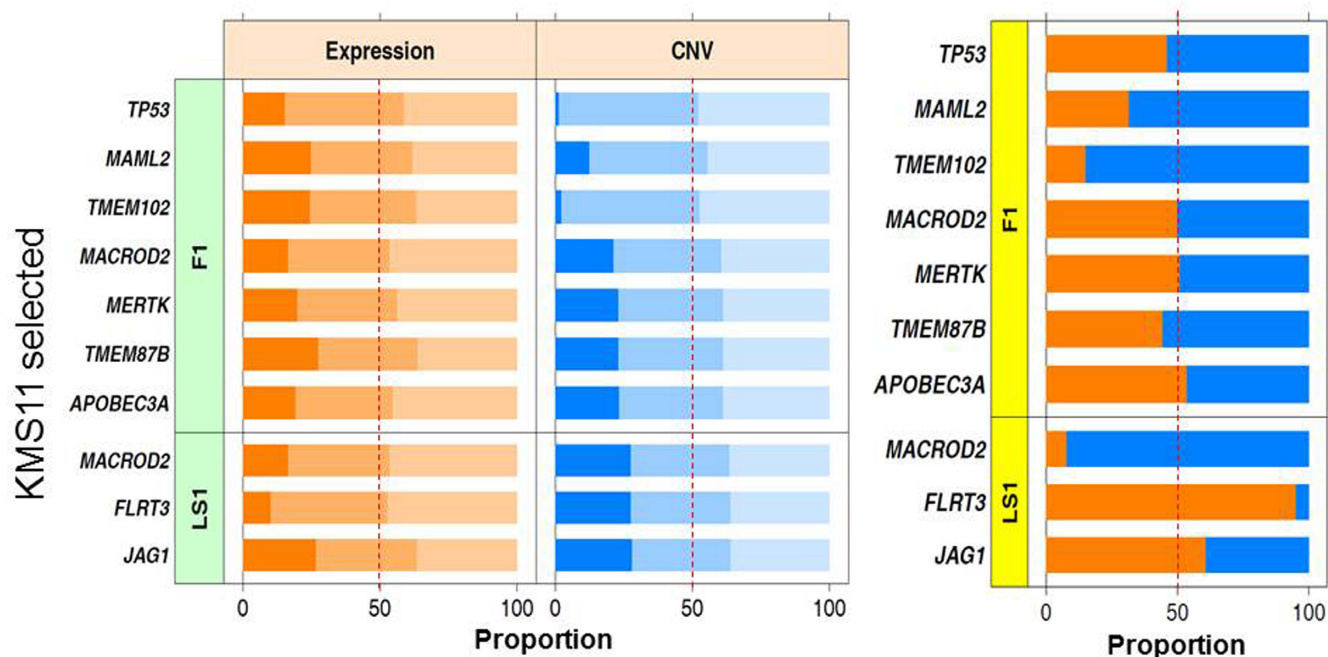


Figure 2. Two feature GISPA identifies KMS11 selected genes with decreased gene expression (GE) and decreased copy number (CN). KMS11 selected, change point 1 (cpt1) gene set results that satisfy the profile of decreased GE with decreased CN by CN segment as focal (F) or large = scale (LS) according to whether the segment's chromosome arm fraction was less than or greater than or equal to 2%, respectively (F1 = focal in cpt1; LS1 = Large-Scale in cpt1). Genes are sorted from the smallest to largest between-feature profile statistic. Left: *Between-cell line differences*. Within each data type: GE (in orange) and CN change (in blue), a stacked bar denoting the percent contribution from each cell line to the summed total of each feature, GE and CN, is displayed along a color gradient from darkest (KMS11) to medium (MM1s) to lightest (RPMI) shades. Among all genes selected, as expected, KMS11 percent contribution to total changes in each feature is the smallest for both GE and CN. Right: *Between-feature Differences*. The percent contribution from each feature to the profile is displayed as a stacked bar. The genes, *TP53*, *MAML2*, *TMEM102* and *TMEM87B* show CN change as the prominent feature driving the profile, whereas *APOBEC3A*, *FLRT3* and *JAG1* show GE, and *MERTK* shows GE and CN as prominent features that equally contribute to the profile. Depending upon the CN segment, *MACROD2* is shown with CN or both CN and GE as prominent features.

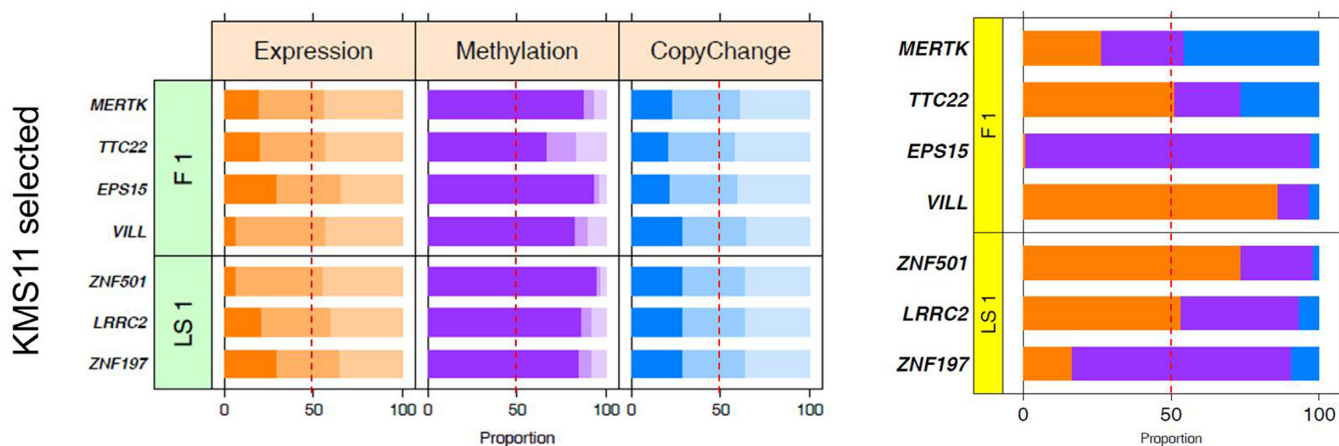


Figure 3. Three Feature GISPA identifies KMS11 selected genes with decreased gene expression (GE), increased CpG methylation, and decreased copy number (CN). KMS11 selected, change point 1 (cpt1) gene results that satisfy the profile of decreased GE with increased CpG methylation and decreased (heterozygous) CN by CN segment as focal (F) or large = scale (LS) according to whether the segment's chromosome arm fraction was less than or greater than or equal to 2%, respectively (F1 = focal in cpt1; LS1 = Large-Scale in cpt1). Genes are sorted from the smallest to largest between-feature profile statistic. Left: *Between-cell line differences*. Within each data type: GE (in orange), CpG methylation (in purple) and CN change (in blue), a stacked bar denoting the percent contribution from each cell line to the summed total of each feature, GE, CN and CpG methylation, is displayed along a color gradient from darkest (KMS11) to medium (MM1s) to lightest (RPMI) shades. Among all genes selected, as expected, KMS11 percent contribution to total changes in each feature is the smallest for GE, largest for CpG methylation and smallest for CN change. Right: *Between-feature Differences*. The percent contribution from each feature to the profile is displayed as a stacked bar. The genes, *EPS15* and *ZNF187* show CpG methylation as prominent features driving the profile, whereas *TTC22*, *VILL*, *ZNF501* and *LRRC2* show GE, and *MERTK*, *CN*.

ticular importance in MM as they contain tumor suppressor genes, *FAF1* and *CDKN2C*; both are in close proximity to *TTC22* and *EPS15*. This raises the possibility that these genes are co-deleted. However, the prominent CpG methylation change of *EPS15* implies that loss of function of *EPS15* may also have functional relevance for MM. We confirmed *EPS15* deletion by FISH and CpG methylation by bisulfite sequencing as specific to the KMS11 cell line (Supplementary Figure S6). Notably, *EPS15* was not identified using an IBI approach, even with a more relaxed *P*-value of 0.05 applied to the results. Among the MM1s and RPMI selected gene set results (Supplementary Figure S5), GISPA identified several known TSG's linked to numerous cancers, including *CDKN2A*, *PARK2*, *ARID1B*, *PTPRD* and *L3MBTL4*, some of which have known associations with MM (17,20–23).

Clinical and biological relevance

Clinical relevance. The large heterogeneity among variants precluded a direct clinical association analysis. In this case, we demonstrate the use of GISPA to define sample-specific gene sets specific to clinically distinct classes that satisfy the two-feature profile of variants with increased GE. We therefore investigated whether the two feature profile of decreased GE and decreased CN cell line specific GISPA gene set (Figure 2) had any clinical relevance. We applied SISPA to Exome-Seq somatic mutations and CN, and RNA-Seq GE from 377 newly diagnosed MM patients enrolled in the coMMpass trial (Appendix 3: Methods) (7) to define pre-treatment samples with similar profiles. SISPA identified 15 samples with decreased CN and decreased GE profile based on the KMS11-specific *cpt1* gene set (Figure 2). These fifteen patients had a significantly ($P = 0.011$) shorter overall survival (OS) as compared to the 362 samples without profile activity (HR = 3.61; 95% CI = (1.24, 10.46); Figure 4B), with no significant differences in progression-free survival (PFS). As a special case, we applied SISPA to define sample classes with decreased CN and GE profiles for each gene in the KMS11-specific *cpt1* gene set. Samples with this profile for the genes, TP53, TMEM102, MACROD2, FLRT3 and JAG1, had significantly shorter OS compared to samples without this profile ($P < 0.05$). The other genes could not be tested due to too few samples with profile activity. Crucially, the HR for the combined gene set was largest in magnitude, suggesting that the observed survival effect was not driven entirely by one gene with this profile, but rather the combined genes had a cumulative effect. This is relevant since deletions of 17p containing TP53 are an indicator of poor prognosis in MM. Additionally, there was no significant association between samples with and without profile activity and the presence of a t(4;14) translocation based on cytogenetic data.

Lack of DNA methylation data in coMMpass precluded direct interrogation of the three-feature profile gene set (Figure 3). However, with the exception of two genes, CN and GE were prominent features driving this profile (Figure 3B). Therefore, we applied SISPA based on the three-feature gene set (Figure 3) to the available, prominent two features (CN down, GE down) mostly associated with this gene set. SISPA identified 25 patients with profile activity (Figure

4A), which compared to the 352 without profile activity, had significantly ($P = 0.027$) shorter PFS (HR = 2.38; 95% CI = 1.07, 5.28; Figure 4B, bottom), and borderline significantly ($P = 0.057$) shorter OS (HR = 2.69, 95% CI = 0.93, 7.82).

Driver mutations. The prognostic relevance of coding mutations in MM is not well understood. The major reason is that with the exception of those in KRAS, NRAS and BRAF, most are relatively infrequent. Thus, popular models such as MutSig cannot inform on driver status. However, an infrequently mutated gene with a functional association to a known cancer gene or pathway may have a driving effect, regardless of mutation frequency. We applied a recent software tool, HyperModules, to find frequently mutated gene modules with clinical associations (9).

Using data from the coMMpass trial, we identified 58 patients with the following extreme prognostic classes: (i) death or relapse within three months of treatment (unfavorable-worse; $n = 13$); (ii) death or relapse at greater than three months of treatment with at least one year of follow-up (unfavorable-bad; $n = 22$); and (iii) alive and no relapse with at least 2 years of follow-up (favorable; $n = 23$). We used GISPA to pre-filter the mutation list passed to HyperModules based on the following hypothesis: an allelic GE skewed toward the mutant implies either loss of the wild type allele through deletion or epigenetic silencing, or amplification/hypomethylation of the mutant allele. We reasoned that the ability to identify such scenarios could increase the likelihood of identifying mutations with driver status. We applied two-feature GISPA to identify somatic mutations with increased GE using 58 samples. A straightforward application of GISPA, in which an average of each data feature is used to define the profile statistic, was not possible as most mutations were found in only one sample, and thus the mean of a mutation type is at or near zero. Instead, we formed 6578 trios using one patient from each of the three prognostic classes, and combined all sample-specific, expressed somatic mutations above *cpt1* into a unified mutation list.

We applied HyperModules to our GISPA gene list using as input a network derived from a random data split of the 58 patients into: (i) $n = 29$ patients ($n = 19$ unfavorable; $n = 10$ favorable prognosis) using 529 GISPA-defined, *cpt1* expressed somatic gene mutations and their identified close connectors from the Reactome database (8); and (ii) 571 GISPA-defined, *cpt1* expressed somatic gene mutations obtained from the other $n = 29$ patients ($n = 16$ unfavorable; $n = 13$ favorable prognosis) and their prognosis as either 'favorable' or 'unfavorable'. We identified a 16 gene module (P -value = 0.10) that included 9 genes, TP53 and TRAF3 among them, with expressed somatic mutations enriched in 10 newly diagnosed MM patients with unfavorable prognosis (Supplementary Figure S7). TP53, a well-known tumor suppressor gene, is also mutated in MM, although only in 3% of cases (24). Activation of the non-canonical NF- κ B pathway by TRAF3 inactivation has been associated with dexamethasone resistance and proteasome inhibitor sensitivity (25). Using this approach, we identified TRAF3 expressed mutations with no corresponding NFKB1 mutation as one of the module connections associated with unfavor-

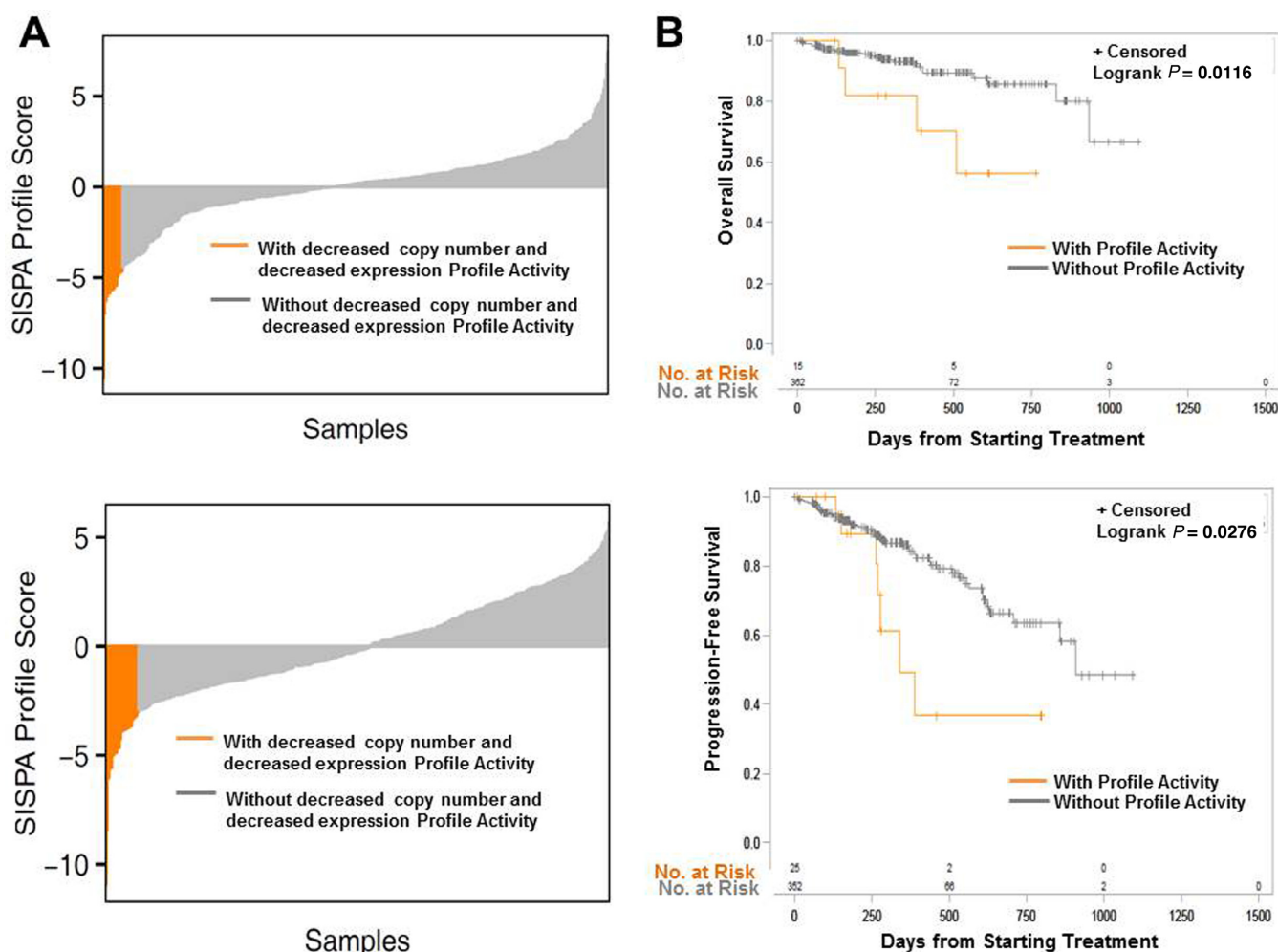


Figure 4. SISPA identifies patient samples clinical relevance of GISPA-defined gene sets. (A) Waterfall plot of SISPA scores based on pre-treatment copy number (CN) and gene expression (GE) data on 377 newly diagnosed MM patient samples from the coMMpass trial using GISPA-defined KMS11 selected, change point 1 gene set results that satisfy the profile of (i) decreased CN with decreased GE (top; Figure 2) or (ii) decreased GE with increased CpG methylation and decreased CN (bottom; Figure 3) to define samples with (in orange) and without (in grey) decreased CN and decreased GE profile activity at pre-treatment. The KMS11 selected, GISPA-defined, three feature (GE, CN, methylation) gene set from Figure 3 was applied to the two features (GE, CN) in the clinical data set from the coMMpass trial, since in the absence of available methylation data, with the exception of two genes, GE and CN were prominent features. (B) Kaplan-Meier overall survival (top) and time to progression (bottom) plots for the SISPA defined sample classes. A log rank test was used to compare survival curves and obtain corresponding *P*-values. The 15 samples with the profile activity (in orange) of decreased GE and decreased CN based on the KMS11 selected gene set corresponding to this same profile (top) have significantly shorter survival as compared to the 362 samples without this profile activity (in grey). The 25 samples (in orange) with profile activity of decreased GE and decreased CN based on the KMS11 selected gene set corresponding to the profile of decreased GE with increased CpG methylation and decreased CN (bottom) have significantly shorter time to progression as compared to the 352 samples without this profile activity (in grey).

able prognosis, revealing a potentially new connection between these genes.

We also identified a significant ($P = 0.05$) 11 gene module that included 7 genes, BRAF among them, with expressed somatic mutations enriched in 8 newly diagnosed MM patients with favorable prognosis (Supplementary Figure S8). In recent years, BRAF-V600E has been shown as a promising 'actionable' mutation that can be targeted for treatment (26). We found the BRAF-G466E mutation as potentially associated with favorable prognosis. This mutation has been previously reported as a low-frequency, cancer-associated variant classified as an impaired activity mutant (less than wild-type BRAF activity) that moderately increases ERK activation (27). Notably, neither prog-

nosis module was identified among the 13 significant ($P < 0.05$) modules when applying HyperModules to the full, GISPA-unfiltered, mutation data from these patients, indicating that the additional information on GE changes from using GISPA had an effect that resulted in a very focused set of modules. The separation of patient prognosis by mutation profiles suggests an underlying common mechanism worthy of separate investigation. The mutation network enriched in the favorable prognosis class in particular is a research area that is aligned with the timely NCI exceptional responder's initiative.

DISCUSSION

Although a comparison of our GISPA results based on different analytic approaches applied to the same data would be ideal, such comparisons are not straightforward within our context for many reasons, including the single sample sizes for each class, the several diverse genome-wide data types simultaneously examined, and the supervised setting. The nature of our experiment is such that three single sample cell lines are to be compared among diverse, genomewide data types in order to characterize each cell line based on a priori specified, relative changes. Most, if not all of the existing ‘integrated’ analysis tools are unable to accommodate single sample analysis, a comparison of more than two classes, and a priori specified change, and more than two data types, for a direct comparison. By introducing other tools that are appropriate for some but not all experimental design aspects would produce ‘ad hoc’ results, and require a further understanding of differences in how the design aspects were handled, in addition to differences in results. The one approach that does offer a more direct comparison with GISPA and is frequently used and straightforward to understand is the integration by intersection. We compared our three-feature GISPA results to those based on an intersecting approach using commonly applied thresholds and fold-change. A comparison based on three features was done to highlight one of the novelties of the GISPA method in being able to accommodate several diverse, genomewide data types. Despite conducting the integration by intersection analysis based on a specific, hypothesized change, by implementing one-sided tests for greater statistical power, no reported genes were selected as in common among the intersecting results and thus, no genes were selected whose change in three features was characteristic of the KMS11 selected cell line; the cell line we highlighted throughout the paper. By comparison, our GISPA approached identified seven such genes. While we have not performed an exhaustive confirmation of all seven genes selected based on our GISPA approach, the results have been supported in part by its ability to identify known genomic changes in the cell lines and by experimental validation of a novel, KMS11-cell line selected gene, EPS15, in terms of decreased methylation, increased expression and decreased copy number relative to multiple myeloma cell lines, MM1s and RPMI-8226.

Our GISPA methodology identified both known and novel gene abnormalities specific to each MM cell line within the context of a priori specified molecular profiles. Furthermore, our analyses show that these gene sets have clinical and biological relevance. While this proof of principle was based on a three sample, three class comparison, the methods can be generalized to n comparisons among n classes with each class containing multiple samples. We further show how GISPA can be used on a larger patient data set with three prognostic classes, each containing several samples per class, to define mutation profiles enriched in each class. Methods to filter mutations as drivers is of considerable interest and remains a challenge, mainly because of lack of approaches to address the low frequency of potential key mutations. Our novel application of GISPA results to HyperModules offered new prognostic associations for

known mutations. There are many potential applications of GISPA, including deriving high, standard and low risk cancer profiles or mutation profiles over a time-course experiment. SISPA could subsequently be used to classify future samples or to identify the best cell line models for each category. Using SISPA to define sample classes with and without profile activity, a Gene Set Enrichment Analysis (28) may then be performed to further identify differential pathways.

While we provided examples of several profiles that may be tested as characterizing each cell line using GISPA, there are limitations to the data types that precludes an exhaustive search. For example, the GISPA-defined, *cpt1* gene set supporting the three feature profile of decreased GE with increased CpG methylation and decreased CN as characteristic of each cell line (Figure 3), may in fact be driven by miRNA-mediated down-regulation. Based on a query of individual genes against several miRNA databases (see Appendix 3: Methods), we examined whether any miRNA commonly targets all or most genes in this set and identified *hsa-miR-656* as predicted to target 6 of the seven genes, followed by *hsa-miR-587*, targeting five of the seven genes (Supplementary Table S5); both predictions included EPS15.

The GISPA approach may be optimized and modified at any step, without loss of generality of the overall method. For example, one may incorporate weights into the WFPS or BFPS profile to penalize poorly supported coding variants or to place a greater emphasis on a particular feature. Additionally, one may estimate the significance of ranked GISPA-defined gene sets (see Appendix, ‘Gene Set Significance’). For example, the three feature profile identified 7 (out of 10) significant GISPA-defined KMS11 selected gene sets with decreased GE and CN and increased methylation.

As newer technology is generated, tools such as GISPA will help to integrate them with existing diverse data types in a way that tests specified profiles as drivers of a phenotype to infer new clinical and biological associations that may be used for their better understanding.

AVAILABILITY

The R code to run GISPA is available for download at <https://github.com/BhaktiDwivedi>. The Bioconductor R package for SISPA is available at <https://www.bioconductor.org/packages/release/bioc/html/SISPA.html>.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Leukemia and Lymphoma Society Translational Research Program Award (to J.K.); Georgia Research Alliance Scientist Award (J.K.); a Team Science Seed Funding from the Winship Cancer Institute of Emory University (L.H.B., S.L., M.R.R.); Biostatistics and Bioinformatics Shared Resource of Winship Cancer Institute of Emory University and NIH/NCI [Award number P30CA138292, in part]. The content is solely the responsibility of the authors and

does not necessarily represent the official views of the NIH. Funding for open access charge: Georgia Research Alliance Scientist Award.

Conflict of interest statement. None declared.

REFERENCES

- Barbie, D.A., Tamayo, P., Boehm, J.S., Kim, S.Y., Moody, S.E., Dunn, I.F., Schinzel, A.C., Sandy, P., Meylan, E., Scholl, C. *et al.* (2009) Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature*, **462**, 108–112.
- Hanzelmann, S., Castelo, R. and Guinney, J. (2013) GSEA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics*, **14**, 7.
- Namba, V.N., Lingjærde, O.C., Russnes, H.G., Volla, H.K., Frigessi, A. and Børresen-Dale, A.L. (2014) Principles and methods of integrative genomic analyses in cancer. *Nat. Rev. Cancer*, **14**, 299–313.
- Namba, M., Ohtsuki, T., Mori, M., Togawa, A., Wada, H., Sugihara, T., Yawata, Y. and Kimoto, T. (1989) Establishment of five human myeloma cell lines. *In Vitro Cellular Dev. Biol.*, **25**, 723–729.
- Moalli, P.A., Pillay, S., Weiner, D., Leikin, R. and Rosen, S.T. (1992) A mechanism of resistance to glucocorticoids in multiple myeloma: transient expression of a truncated glucocorticoid receptor mRNA. *Blood*, **79**, 213–222.
- Matsuoka, Y., Moore, G.E., Yagi, Y. and Pressman, D. (1967) Production of free light chains of immunoglobulin by a hematopoietic cell line derived from a patient with multiple myeloma. *Proc. Soc. Exp. Biol. Med.*, **125**, 1246–1250.
- Lonial, S., Yellapantula, V.D., Liang, W., Kurdoglu, A., Aldrich, J., Legendre, C.M., Stephenson, K., Adkins, J., McDonald, J., Helland, A. *et al.* (2014) Interim analysis of the MmrF Compass Trial: identification of novel rearrangements potentially associated with disease initiation and progression. *ASH (Annual Meeting Abstracts)*, (Abstract 722).
- Wu, G., Dawson, E., Duong, A., Haw, R. and Stein, L. (2014) ReactomeFIViz: a Cytoscape app for pathway and network-based data analysis. *F1000Research*, **3**, 146.
- Leung, A., Bader, G.D. and Reimand, J. (2014) HyperModules: identifying clinically and phenotypically significant network modules with disease mutations for biomarker discovery. *Bioinformatics*, **30**, 2230–2232.
- Bindea, G., Galon, J. and Mlecnik, B. (2013) CluePediaCytoscape plugin: pathway insights using integrated experimental and in silico data. *Bioinformatics (Oxford, England)*, **29**, 661–663.
- Yang, R., Chen, L., Newman, S., Gandhi, K., Doho, G., Moreno, C.S., Vertino, P.M., Bernal-Mizarchi, L., Lonial, S., Boise, L.H. *et al.* (2014) Integrated analysis of whole-genome paired-end and mate-pair sequencing data for identifying genomic structural variations in multiple myeloma. *Cancer Inform.*, **13**, 49–53.
- Chesi, M., Nardini, E., Brents, L.A., Schröck, E., Ried, T., Kuehl, W.M. and Bergsagel, P.L. (1997) Frequent translocation t(4;14)(p16.3;q32.3) in multiple myeloma is associated with increased expression and activating mutations of fibroblast growth factor receptor 3. *Nat. Genet.*, **16**, 260–264.
- Anders, S. and Huber, W. (2010) Differential expression analysis for sequence count data. *Genome biology*, **11**, R106.
- Cummings, C.T., Deryckere, D., Earp, H.S. and Graham, D.K. (2013) Molecular pathways: MERTK signaling in cancer. *Clin. Cancer Res.*, **19**, 5275–5280.
- Huchtagowder, V., Meyer, R., Mullins, C., Nagarajan, R., DiPersio, J.F., Vij, R., Tomasson, M.H. and Kulkarni, S. (2012) Resequencing analysis of the candidate tyrosine kinase and RAS pathway gene families in multiple myeloma. *Cancer Genet.*, **205**, 474–478.
- Sagawa, M., Shimizu, T., Shimizu, T., Awaya, N., Mitsuhashi, T., Ikeda, Y., Okamoto, S. and Kizaki, M. (2006) Establishment of a new human acute monocytic leukemia cell line TZ-1 with t(1;11)(p32;q23) and fusion gene MLL-EP515. *Leukemia*, **20**, 1566–1571.
- Walker, B.A., Leone, P.E., Chiecchio, L., Dickens, N.J., Jenner, M.W., Boyd, K.D., Johnson, D.C., Gonzalez, D., Dagrada, G.P., Protheroe, R.K. *et al.* (2010) A compendium of myeloma-associated chromosomal copy number abnormalities and their prognostic value. *Blood*, **116**, e56–e65.
- Boyd, K.D., Ross, F.M., Walker, B.A., Wardell, C.P., Tapper, W.J., Chiecchio, L., Dagrada, G., Konn, Z.J., Gregory, W.M., Jackson, G.H. *et al.* (2011) Mapping of chromosome 1p deletions in myeloma identifies FAM46C at 1p12 and CDKN2C at 1p32.3 as being genes in regions associated with adverse survival. *Clin. Cancer Res.*, **17**, 7776–7784.
- Chang, H., Jiang, A., Qi, C., Trieu, Y., Chen, C. and Reece, D. (2010) Impact of genomic aberrations including chromosome 1 abnormalities on the outcome of patients with relapsed or refractory multiple myeloma treated with lenalidomide and dexamethasone. *Leuk. Lymphoma*, **51**, 2084–2091.
- Elnenaie, M.O., Gruszka-Westwood, A.M., A'Hernt, R., Matutes, E., Sirohi, B., Powles, R. and Catovsky, D. (2003) Gene abnormalities in multiple myeloma; the relevance of TP53, MDM2, and CDKN2A. *Haematologica*, **88**, 529–537.
- Gonzalez-Paz, N., Chng, W.J., McClure, R.F., Blood, E., Oken, M.M., Van Ness, B., James, C.D., Kurtin, P.J., Henderson, K., Ahmann, G.J. *et al.* (2007) Tumor suppressor p16 methylation in multiple myeloma: biological and clinical implications. *Blood*, **109**, 1228–1232.
- Walia, V., Prickett, T.D., Kim, J.S., Gartner, J.J., Lin, J.C., Zhou, M., Rosenberg, S.A., Elble, R.C., Solomon, D.A., Waldman, T. *et al.* (2014) Mutational and functional analysis of the tumor-suppressor PTPRD in human melanoma. *Human Mutat.*, **35**, 1301–1310.
- Addou-Klouche, L., Adélaide, J., Finetti, P., Cervera, N., Ferrari, A., Bekhouche, I., Sircoulomb, F., Sotiriou, C., Viens, P., Mouleschoul, S. *et al.* (2010) Loss, mutation and deregulation of L3MBTL4 in breast cancers. *Mol. Cancer*, **9**, 213.
- Chng, W.J., Price-Troska, T., Gonzalez-Paz, N., Van Wier, S., Jacobus, S., Blood, E., Henderson, K., Oken, M., Van Ness, B., Greipp, P. *et al.* (2007) Clinical significance of TP53 mutation in myeloma. *Leukemia*, **21**, 582–584.
- Keats, J.J., Fonseca, R., Chesi, M., Schop, R., Baker, A., Chng, W.J., Van Wier, S., Tiedemann, R., Shi, C.X., Sebag, M. *et al.* (2007) Promiscuous mutations activate the noncanonical NF-kappaB pathway in multiple myeloma. *Cancer Cell*, **12**, 131–144.
- Andrulis, M., Lehnert, N., Capper, D., Penzel, R., Heining, C., Huellein, J., Zenz, T., von Deimling, A., Schirmacher, P., Ho, A.D. *et al.* (2013) Targeting the BRAF V600E mutation in multiple myeloma. *Cancer Discov.*, **3**, 862–869.
- Wan, P.T., Garnett, M.J., Roe, S.M., Lee, S., Niculescu-Duvaz, D., Good, V.M., Jones, C.M., Marshall, C.J., Springer, C.J., Barford, D. *et al.* (2004) Mechanism of activation of the RAF-ERK signaling pathway by oncogenic mutations of B-RAF. *Cell*, **116**, 855–867.
- Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 15545–15550.